Mixture models for phylogenetic analysis in IQ-TREE2

Thomas Wong

School of Computing, Australian National University

Algorithmic Advances and Implementation Challenges:
Developing Practical Tools for Phylogenetic Inference (Nov 18 - 22, 2024)

Model-based Phylogenetics

DNA/Amino acid Sequence Alignment

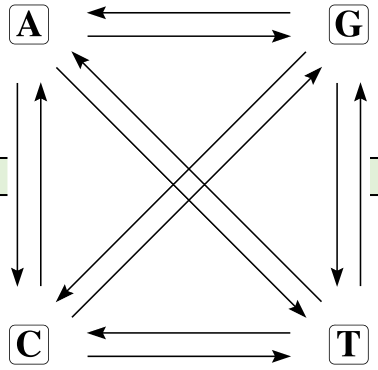
Bird	C	A	A	-	-	-	A	A	T	A
Crocodile	C	A	C	A	–	T	A	C	–	T
Turtle	C	A	C	T	A	T	A	A	G	T
Human	C	A	C	-	-	-	A	C	A	A

Assumptions are:

1. Single model
2. Single tree
3. Sites evolve independently
4. ....

Markov model

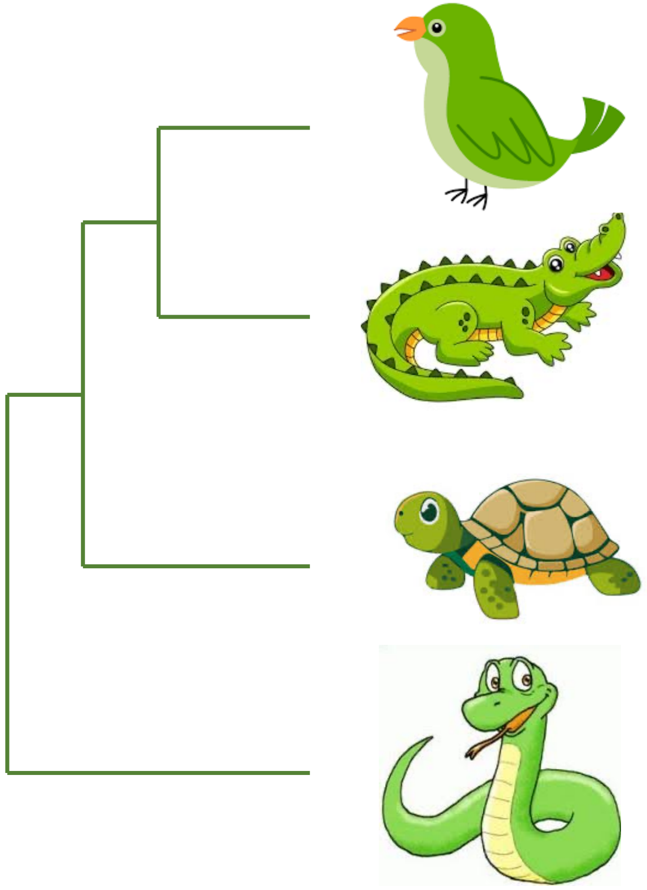
Substitution matrix:



Frequency array:

π_A π_C π_G π_T

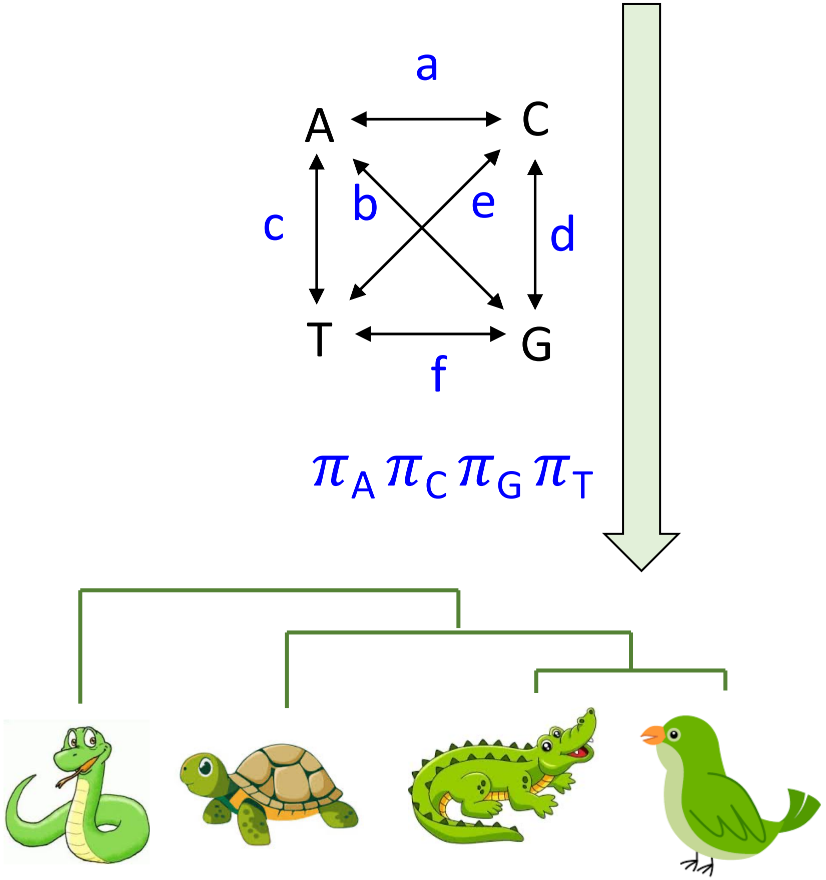
Phylogenetic Tree



Assumption

All alignment sites evolve equally under the same phylogenetic process?

Bird	C	A	A	-	-	-	A	A	T	A
Crocodile	C	A	C	A	-	T	A	C	-	T
Turtle	C	A	C	T	A	T	A	A	G	T
Human	C	A	C	-	-	-	A	C	A	A



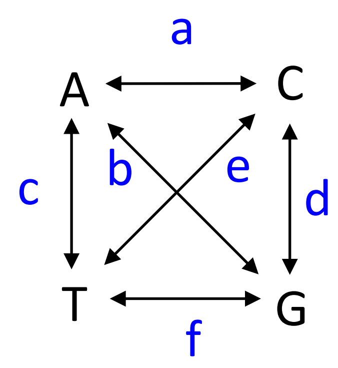
Single model often not true

For example:

- Multiple genes
- On different chromosomes
- Encode different proteins
- Proteins have very different functions
- Under different biological constraints

➢ One model is not good enough to describe the evolutionary process along the alignment

	Gene A					Gene B				
Bird	C	A	A	-	-	-	A	A	T	A
Crocodile	C	A	C	A	-	T	A	C	-	T
Turtle	C	A	C	T	A	T	A	A	G	T
Human	C	A	C	-	-	-	A	C	A	A



$\pi_A \pi_C \pi_G \pi_T$

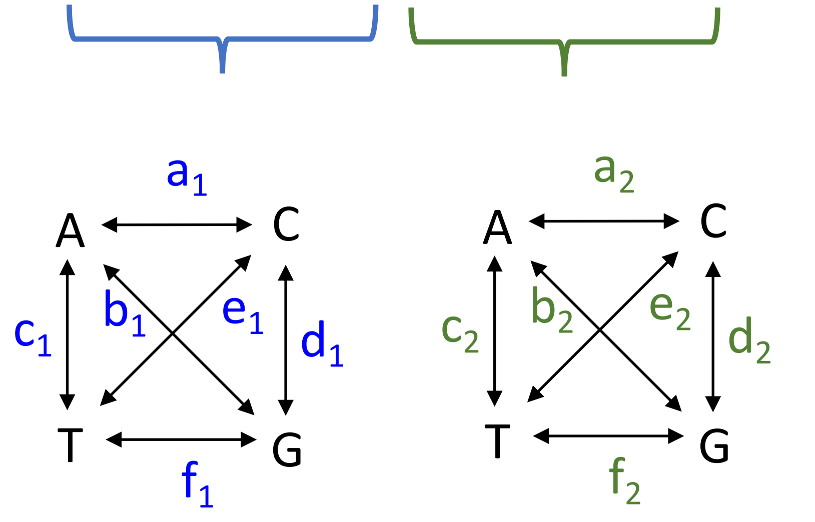
Partition model

- Each partition has a separate model

➢ Often it can fit the data much better than a single model

➢ However,
  ➢ Gene information is absent
  ➢ Gene boundary is not accurate
  ➢ For protein, partition according to protein domain?

	Gene A					Gene B				
Bird	C	A	A	-	-	-	A	A	T	A
Crocodile	C	A	C	A	-	T	A	C	-	T
Turtle	C	A	C	T	A	T	A	A	G	T
Human	C	A	C	-	-	-	A	C	A	A



$\pi_{A1}\, \pi_{C1}\, \pi_{G1}\, \pi_{T1}$ $\pi_{A2}\, \pi_{C2}\, \pi_{G2}\, \pi_{T2}$

Mixture model

It **does not need** any partition information

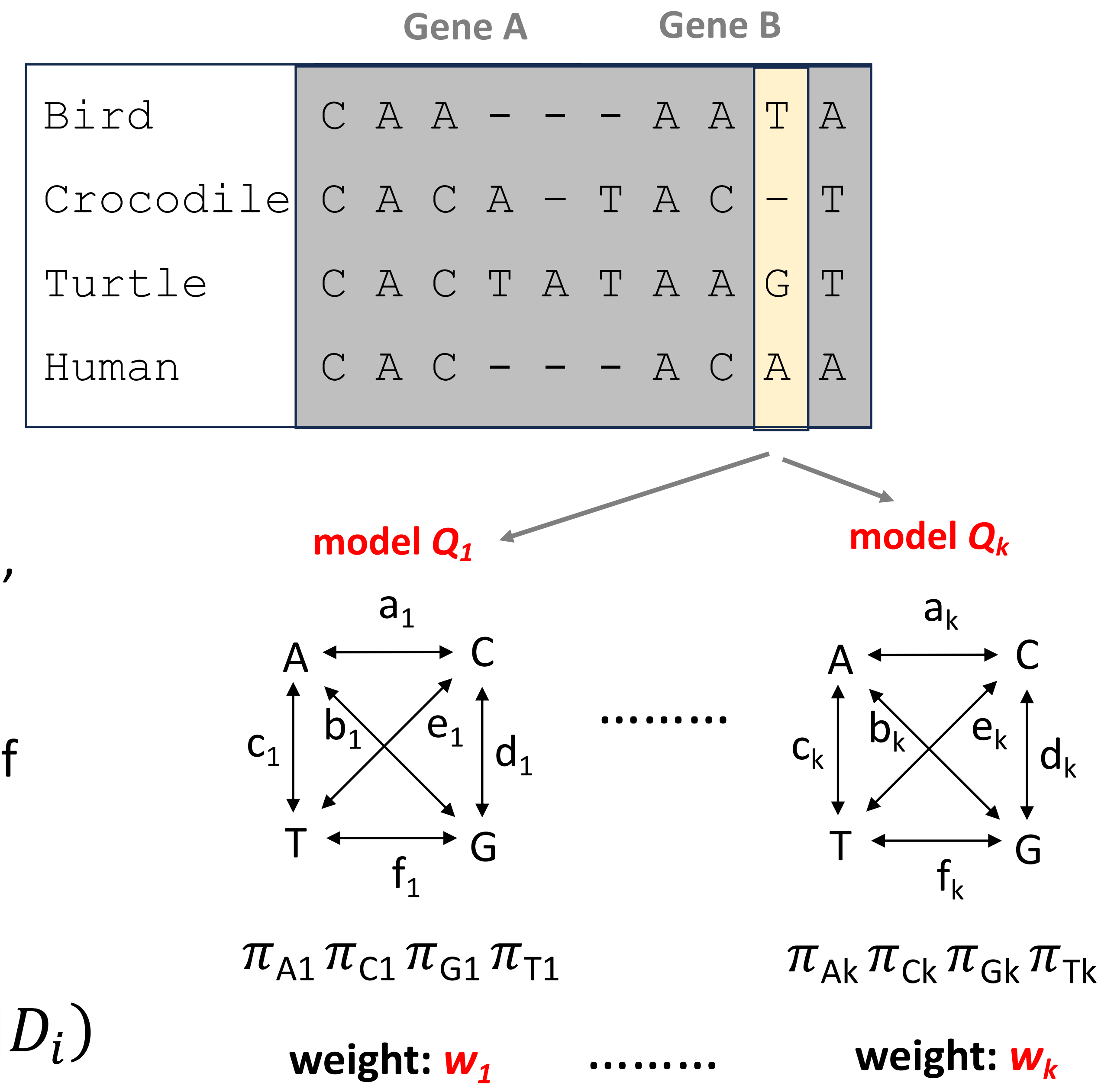
Assume each site evolves under a mixture of ***k-class*** models

Each class can have a **different** model
For example: Q_1 – GTR; Q_2 – HKY; Q_3 – GTR’

Likelihood of each site is a **weighted sum** of likelihoods across all the models

For example, likelihood of site ***i***:

$$L_i = \mathbf{w_1} L(Q_1|D_i) + \cdots + \mathbf{w_k} L(Q_2|D_i)$$



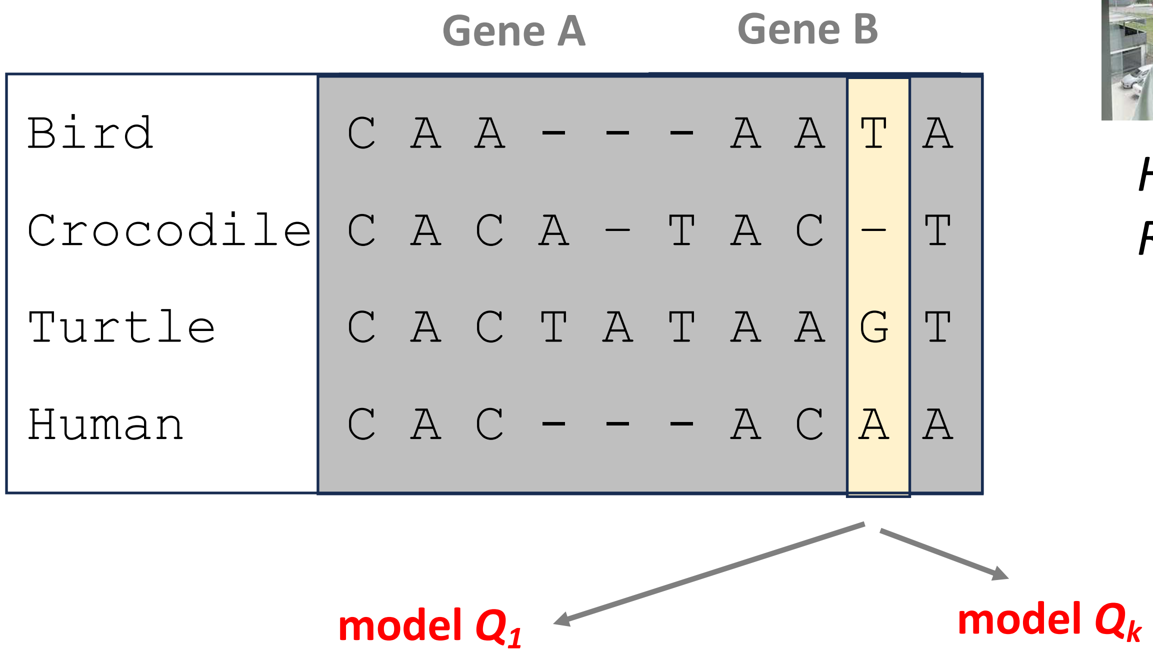
Bird	C	A	A	-	-	-	A	A	T	A
Crocodile	C	A	C	A	–	T	A	C	–	T
Turtle	C	A	C	T	A	T	A	A	G	T
Human	C	A	C	-	-	-	A	C	A	A

Mixture model

Huaiyan Ren

Has been implemented in IQ-TREE for some years

However, this model is not popular



Bird	C	A	A	-	-	-	A	A	T	A
Crocodile	C	A	C	A	-	T	A	C	-	T
Turtle	C	A	C	T	A	T	A	A	G	T
Human	C	A	C	-	-	-	A	C	A	A

Not easy to decide

- the number of classes (i.e. the value of ***k***)?
- which model for each class?

K80

GTR

HKY

- ➢ Introduce a new algorithm MixtureFinder
- ➢ Automatically estimate
  - a. Optimal number of classes
  - b. Optimal model for each class

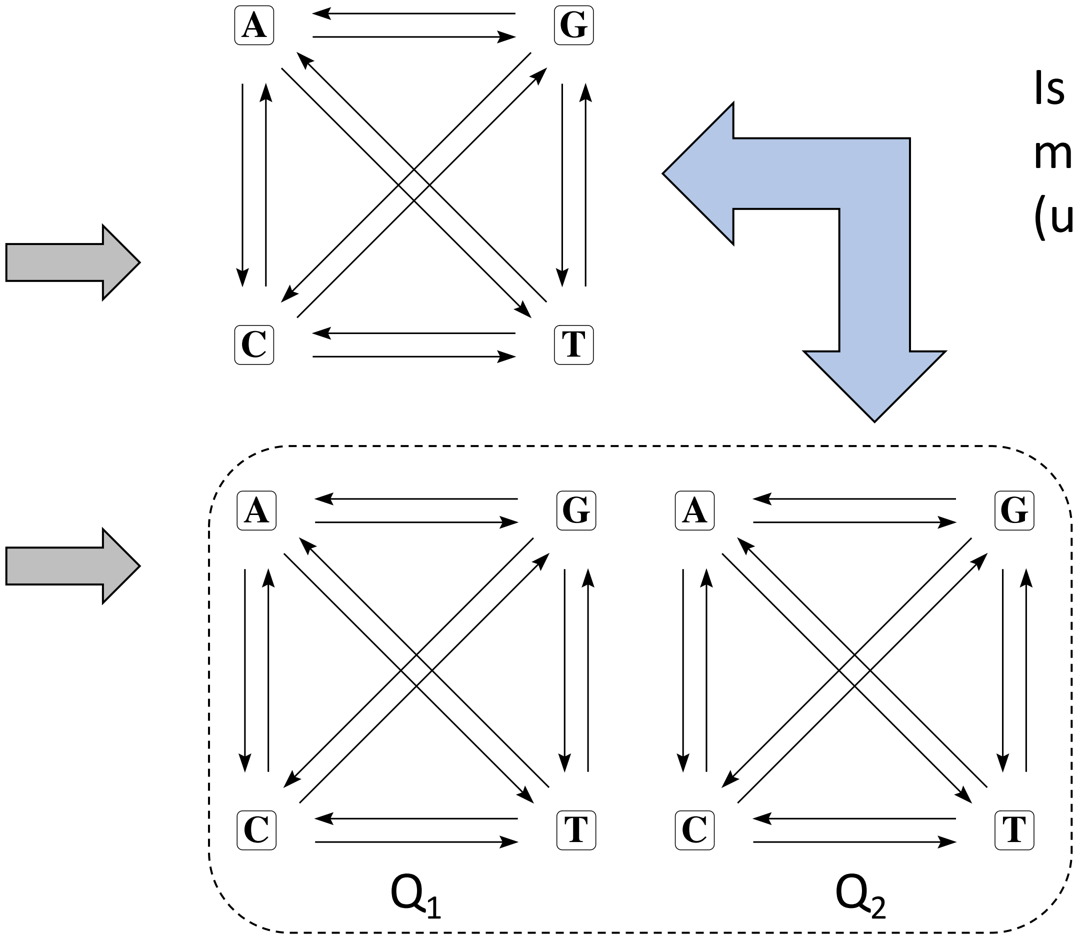
Is a single model good enough?

Examine 19,834 DNA partitions

Sequence Alignment

```
Bird      C A A - - - A A T A
Crocodile C A C A - T A C - T
Turtle    C A C T A T A A G T
Human     C A C - - - A C A A
```

Markov model



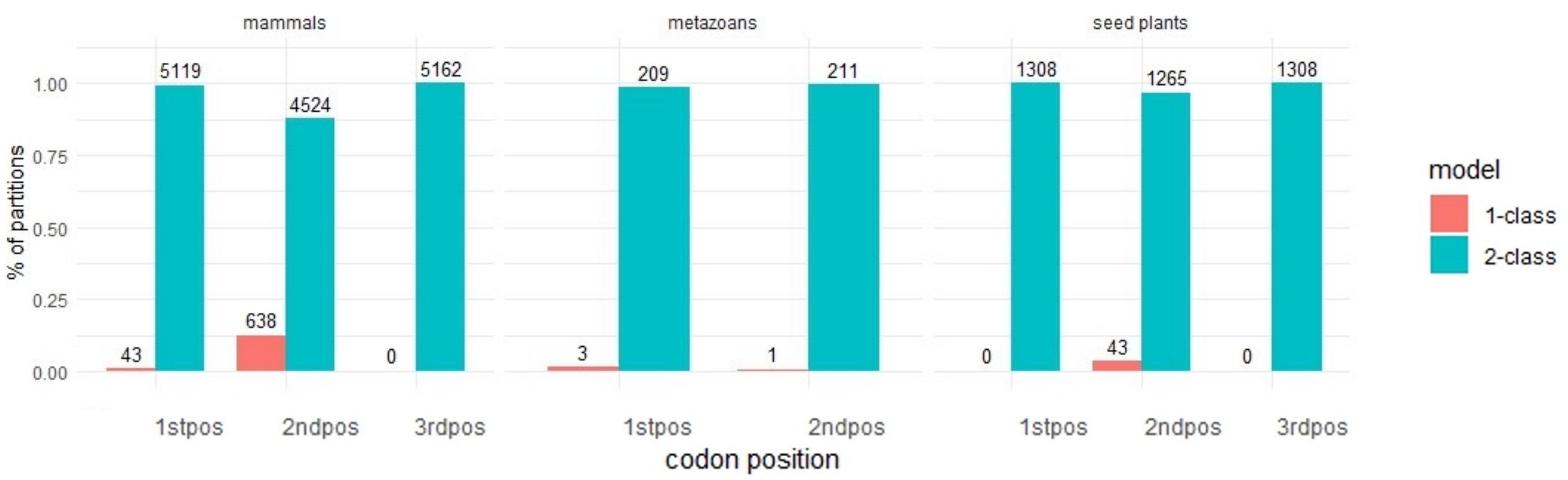
Is mixture model better? (using BIC)

Mixture model of Q_1 and Q_2

Is one model enough?

Results

Among **19,834** DNA partitions, in **19,106 (96.3%)** of them the **2-class** models have **better** BIC value.



MixtureFinder (for DNA)

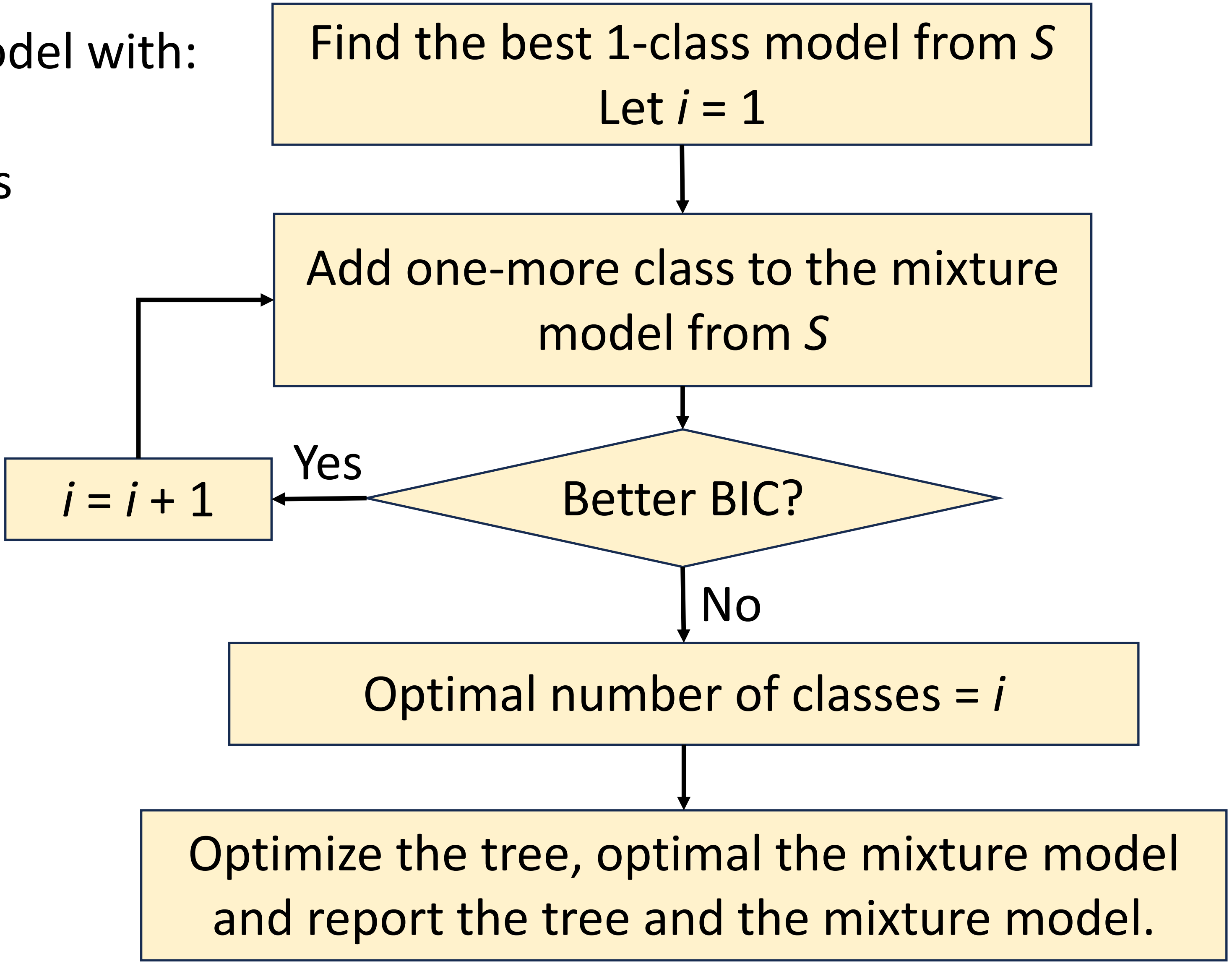
To compute the optimal mixture model with:

a. Optimal number of classes
b. Optimal model for each class

Let *S* be all DNA models to consider.

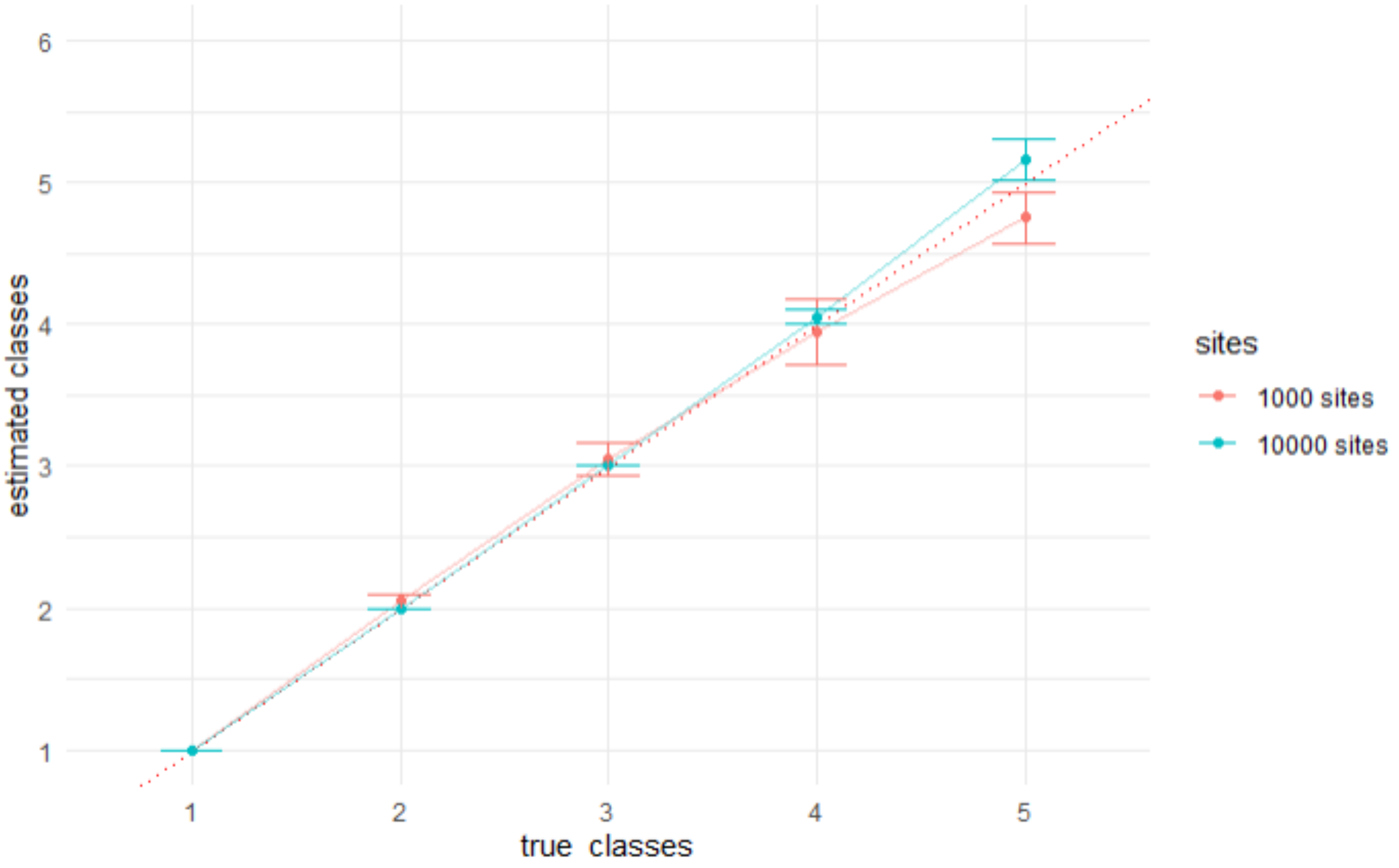
S = {JC, F81, K80, HKY, …, GTR}

Workflow


Find the best 1-class model from *S*
Let *i* = 1
Add one-more class to the mixture model from *S*
Better BIC?
Yes
i = *i* + 1
No
Optimal number of classes = *i*
Optimize the tree, optimal the mixture model and report the tree and the mixture model.

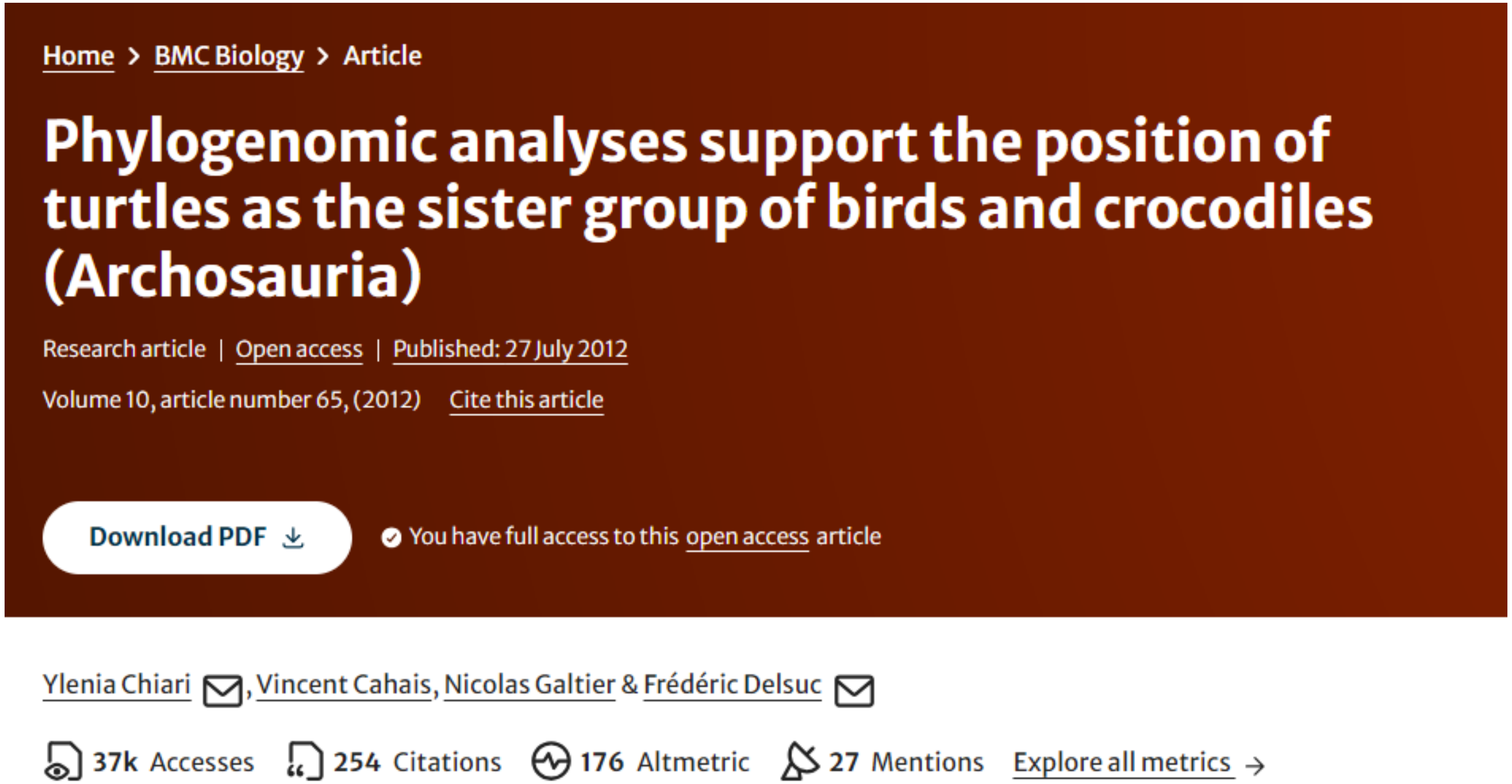
Evaluating MixtureFinder on Simulated data

Estimated number of classes vs simulated number of classes



Experiment 3

Apply the MixtureFinder on the empirical data:



Home > BMC Biology > Article

Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria)

Research article | Open access | Published: 27 July 2012

Volume 10, article number 65, (2012) Cite this article

You have full access to this open access article

Ylenia Chiari, Vincent Cahais, Nicolas Galtier & Frédéric Delsuc

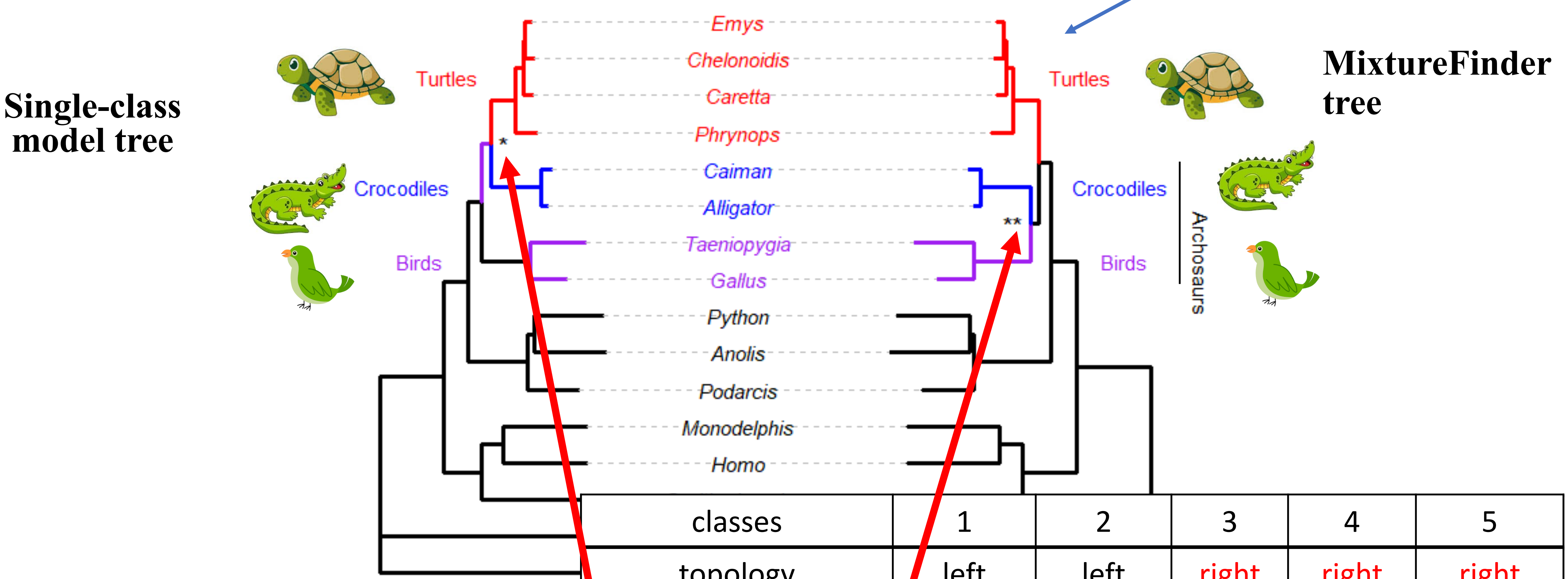
37k Accesses 254 Citations 176 Altmetric 27 Mentions Explore all metrics →

16 vertebrate taxa,
248 genes (187k sites)

Apply the MixtureFinder to the concatenated alignment

Experiment 3

Apply the MixtureFinder on the empirical data:



This is the published tree

classes	1	2	3	4	5
topology	left	left	right	right	right
Bootstrap of node *	71	52	24	5	1
Bootstrap of node **	29	48	86	95	99

Conclusion

- A single substitution model is always not adequate to empirical data set.

- Model is matter! MixtureFinder can recover a better tree than a single model.

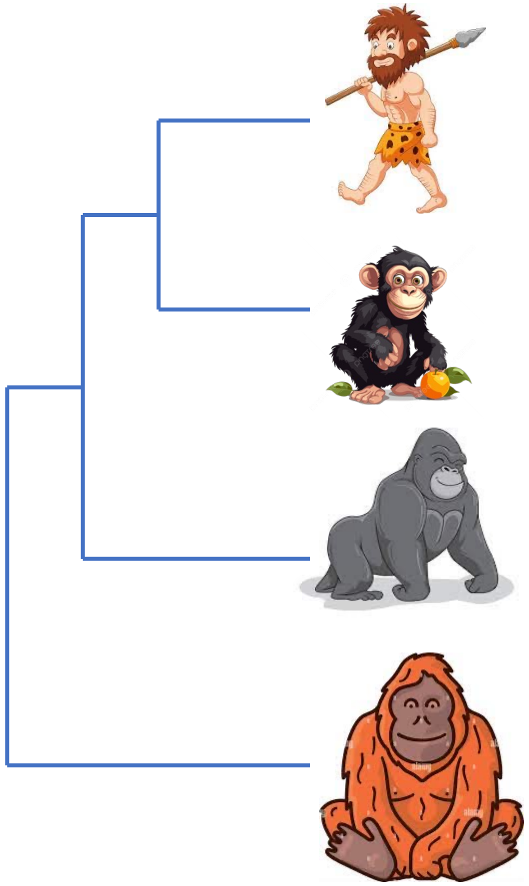
- MixtureFinder is available in IQ-TREE2.

- MixtureFinder now only supports DNA. We may extend it to amino acid in the future.

Preprint: https://www.biorxiv.org/content/10.1101/2024.03.20.586035v2

MAST model, a multi-tree mixture model

Traditional phylogenetic analyses always make this assumption:



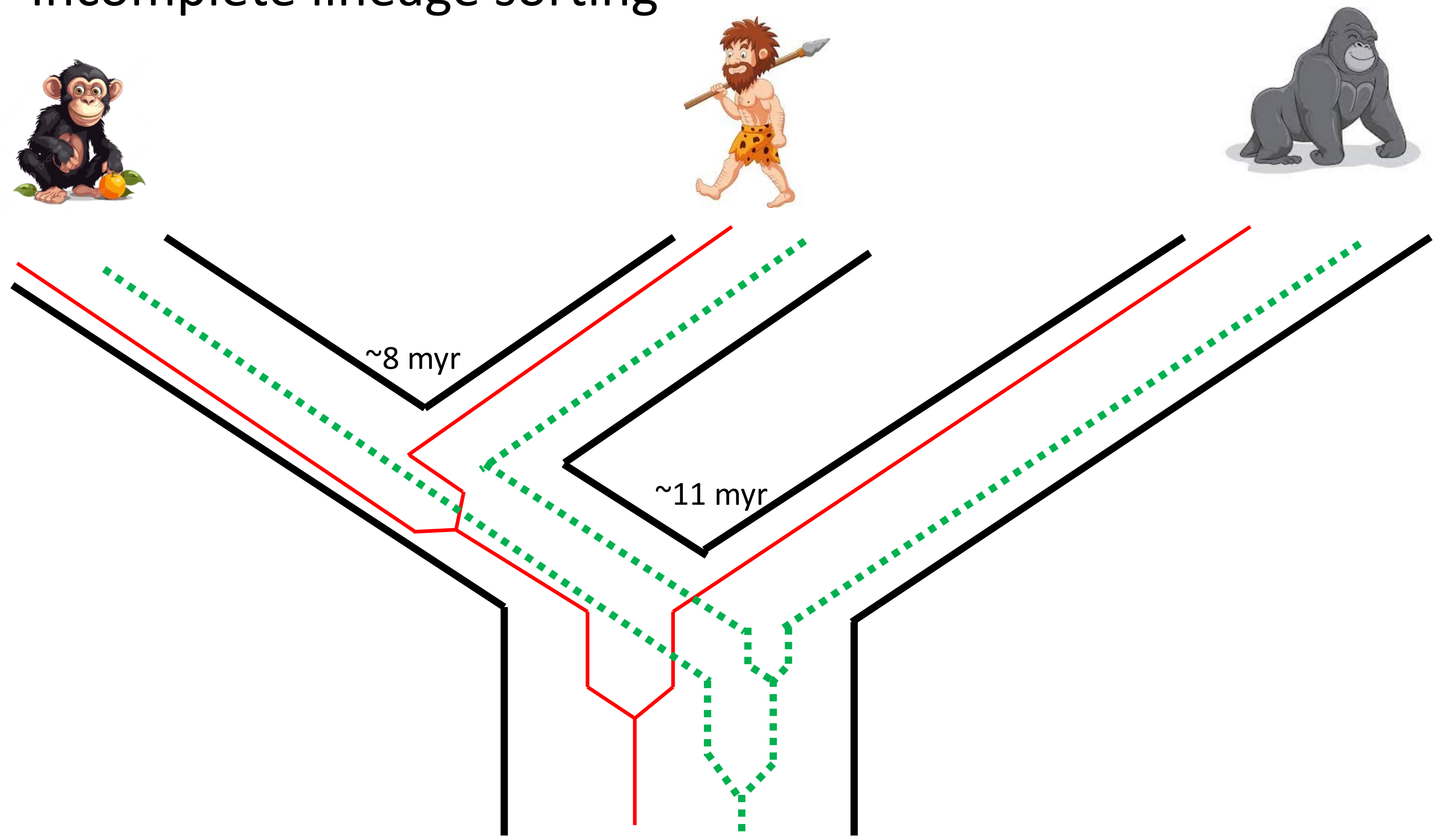
Gene A							Gene B						
C	A	A	-	-	-	A	A	T	A	T	T	A	C
C	A	C	A	-	T	A	C	-	T	T	T	A	C
C	A	C	T	A	T	A	A	G	T	T	T	A	C
C	A	C	-	-	-	A	C	A	A	A	T	A	C

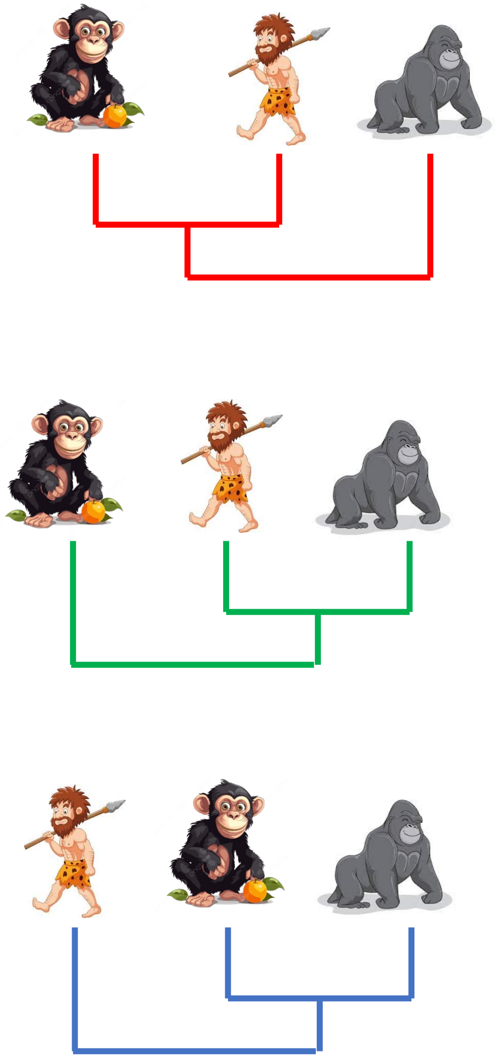
All sites evolved under a single tree

Is it always true?

For example:

Incomplete lineage sorting





Biological processes:

Incomplete lineage sorting

Introgression

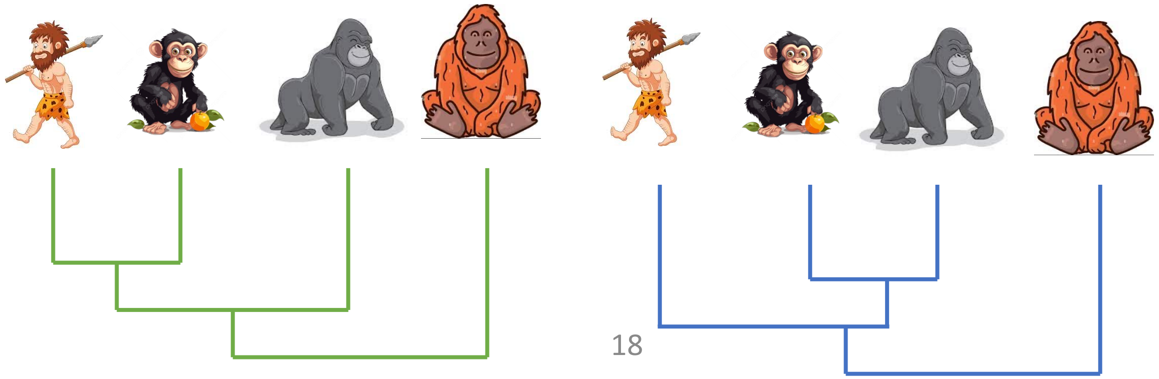
Recombination

Different genomic loci may have evolved under different trees

S1:	A	A	A	-	T	A	A	A	T	T	A	C
S2:	C	T	A	A	C	C	T	T	T	T	A	C
S3:	C	T	A	T	A	A	G	T	T	T	T	A
S4:	C	A	C	-	A	C	A	A	A	T	A	C

Gene A

Gene B



Existing approaches

- Multi-species coalescent model
- Phylogenetic network

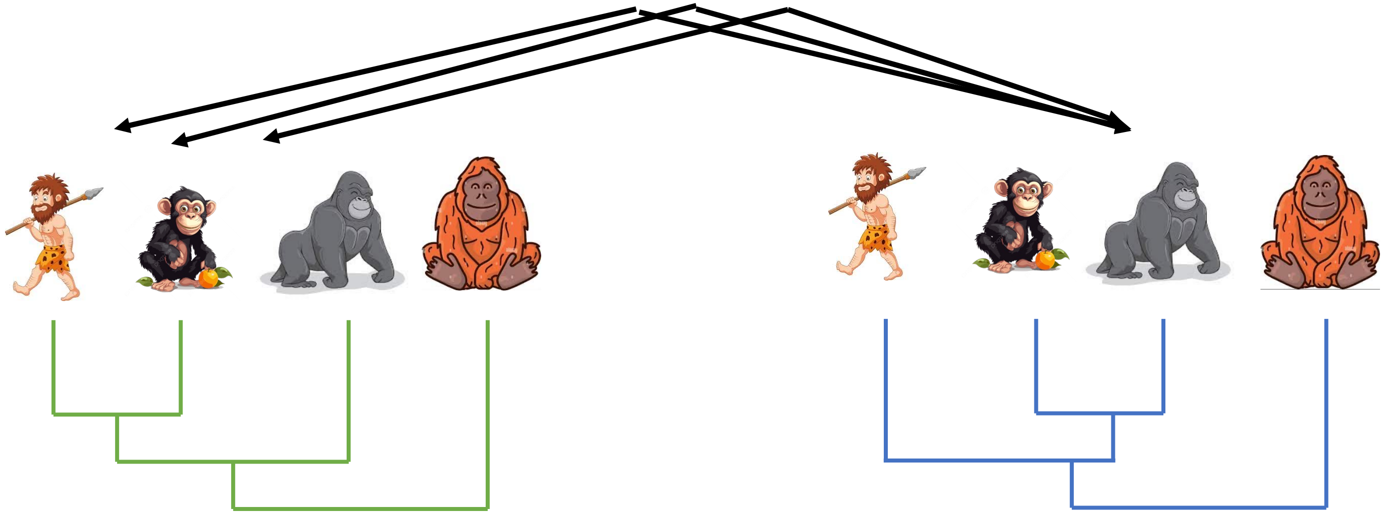
We suggest a new approach:
a Mixture Across Sites and Trees (MAST) model

MAST model

```
S1: A A A - T A A A T T A C
S2: C T A A C C T T T T A C
S3: C T A T A A G T T T T A
S4: C A C - A C A A A T A C
```

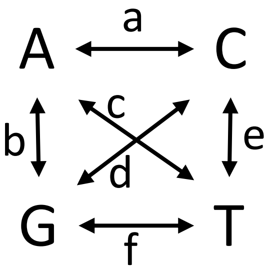
Gene A

Gene B



Assume each site evolved under a mixture of trees

Substitution model



GTR

$a \neq b \neq c \neq d \neq e \neq f$

JC

$a = b = c = d = e = f$

RHAS model

GTR + G

JC + I

MAST model

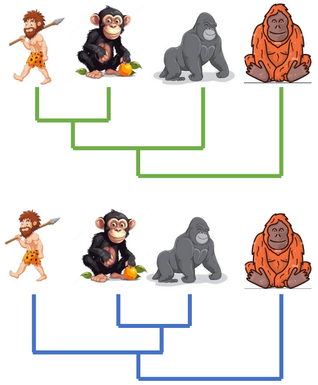
Concatenated alignment

S1:	A	A	-	T	A	A	A	T
S2:	T	A	A	C	C	T	T	T
S3:	T	A	T	A	A	G	T	T
S4:	A	C	-	A	C	A	A	A

Gene A Gene B

$L_{1,1}$ $L_{2,1}$ $L_{3,1}$ $L_{4,1}$ $L_{5,1}$ $L_{6,1}$ $L_{7,1}$ $L_{8,1}$

$L_{1,2}$ $L_{2,2}$ $L_{3,2}$ $L_{4,2}$ $L_{5,2}$ $L_{6,2}$ $L_{7,2}$ $L_{8,2}$



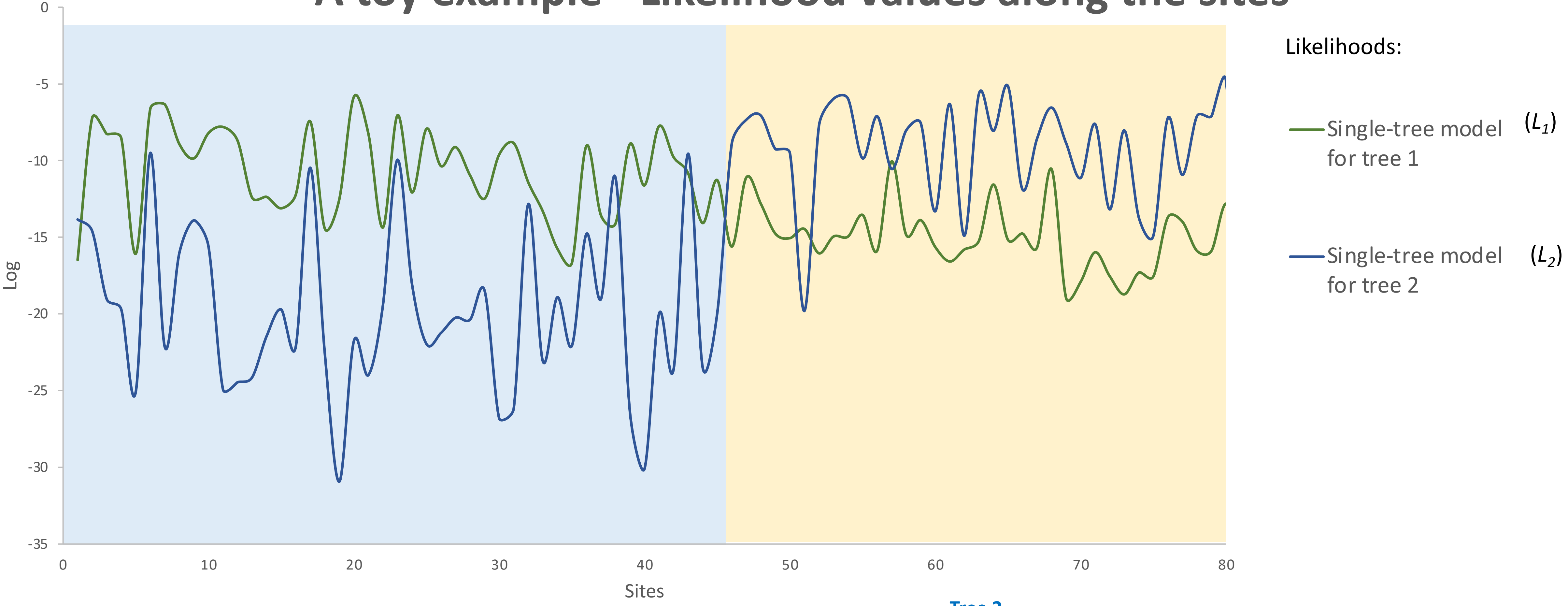
Weight of each tree

w_1

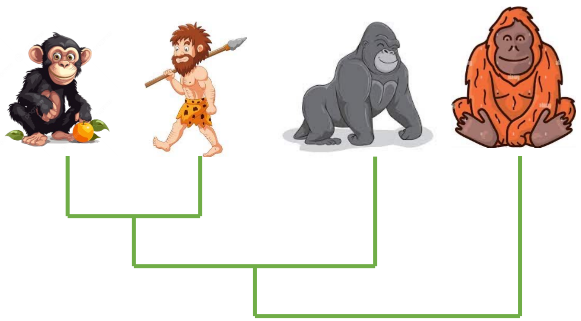
w_2

Likelihood of site *i* (L_i) = $w_1 L_{i,1} + w_2 L_{i,2}$

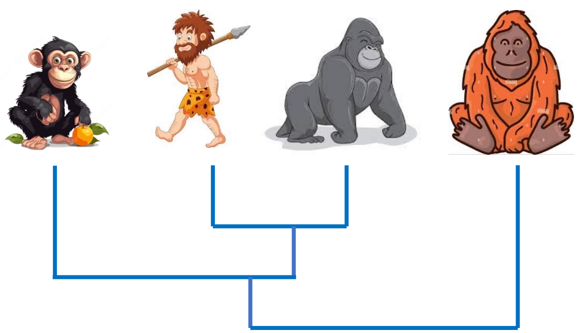
A toy example - Likelihood values along the sites



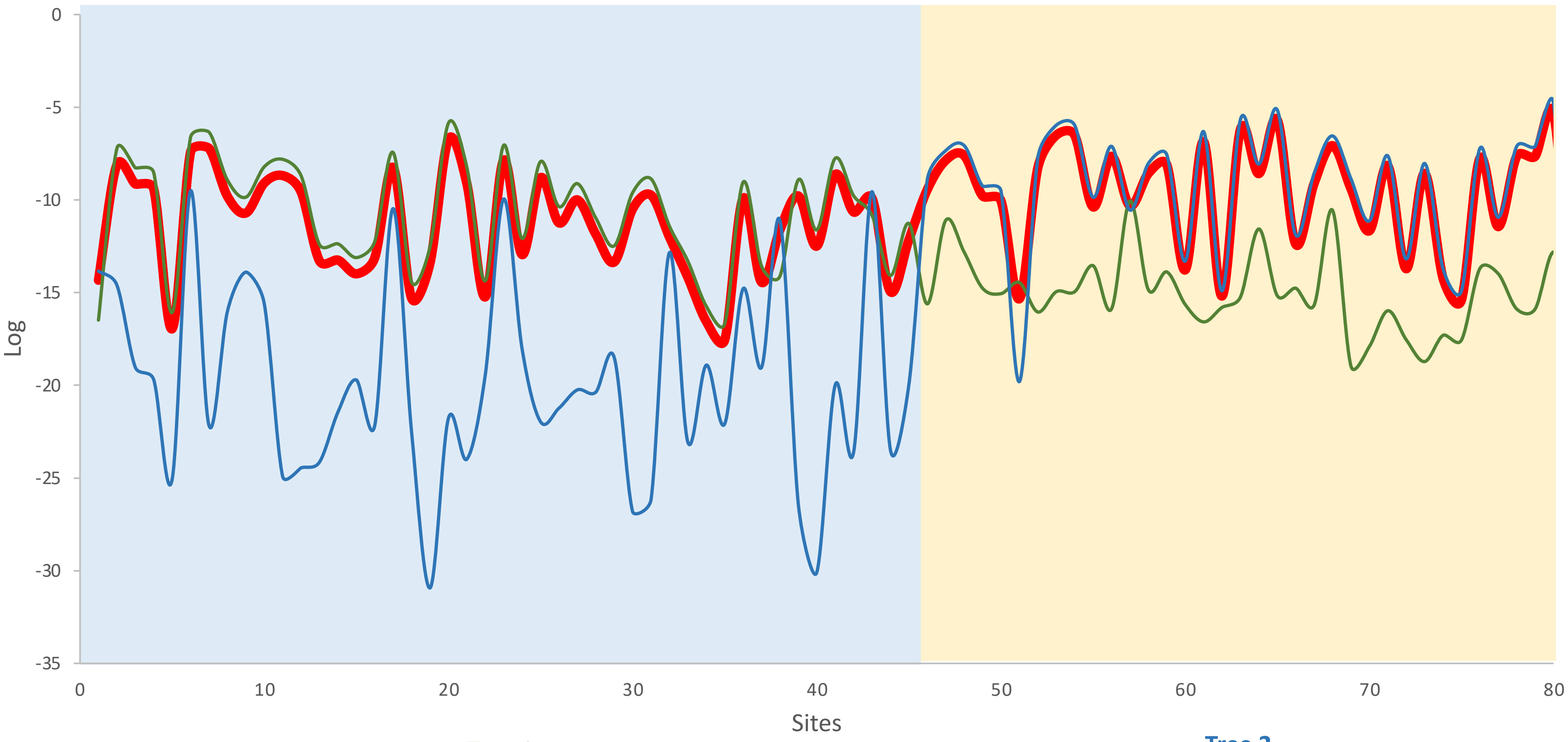
Tree 1



Tree 2



A toy example - Likelihood values along the sites



Likelihoods:

MAST with both trees as an input (L_{MAST})

Single-tree model for tree 1 (L_1)

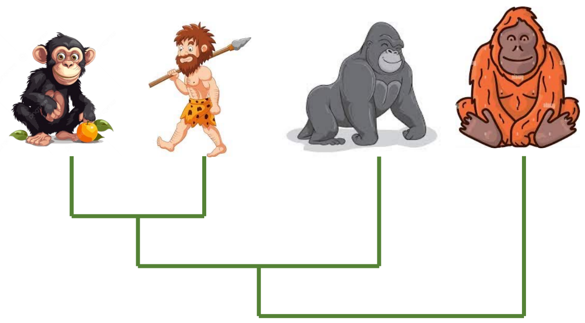
Single-tree model for tree 2 (L_2)

$$L_{MAST} = w_1 L_1 + w_2 L_2$$

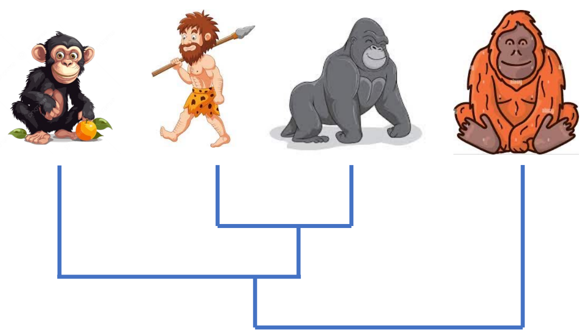
w_1: weight of tree 1

w_2: weight of tree 2

Tree 1



Tree 2



- The MAST model has been implemented in IQ-TREE
- We’ve done lots of simulations to verify its correctness.
- Tutorial: http://www.iqtree.org/doc/Complex-Models#multitree-models

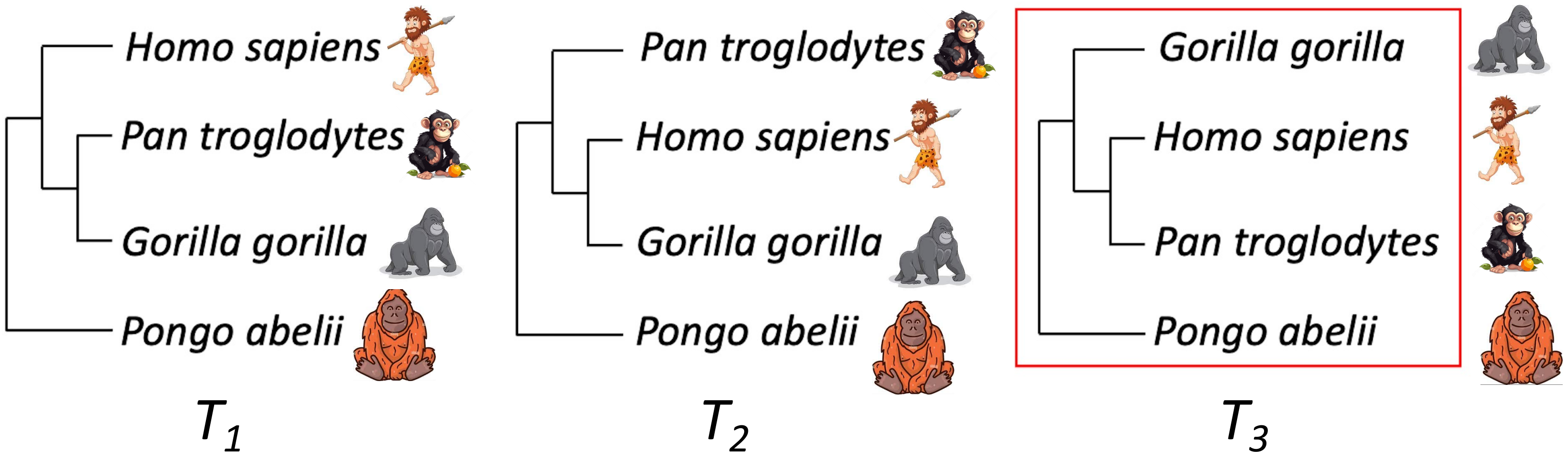
Three empirical experiments

Apply to Human-Chimp-Gorilla data

Number of genes: 1,595

Total length: 1,618,506

- Well-studied four-taxon grouping of human, chimpanzee, gorilla, and orangutan
- The accepted species tree: T_3



Gene tree frequencies:	19.8%	20.1%	60.1%
Our result **Best-fit MAST model weights:**	**17.9%**	**17.4%**	**64.7%**

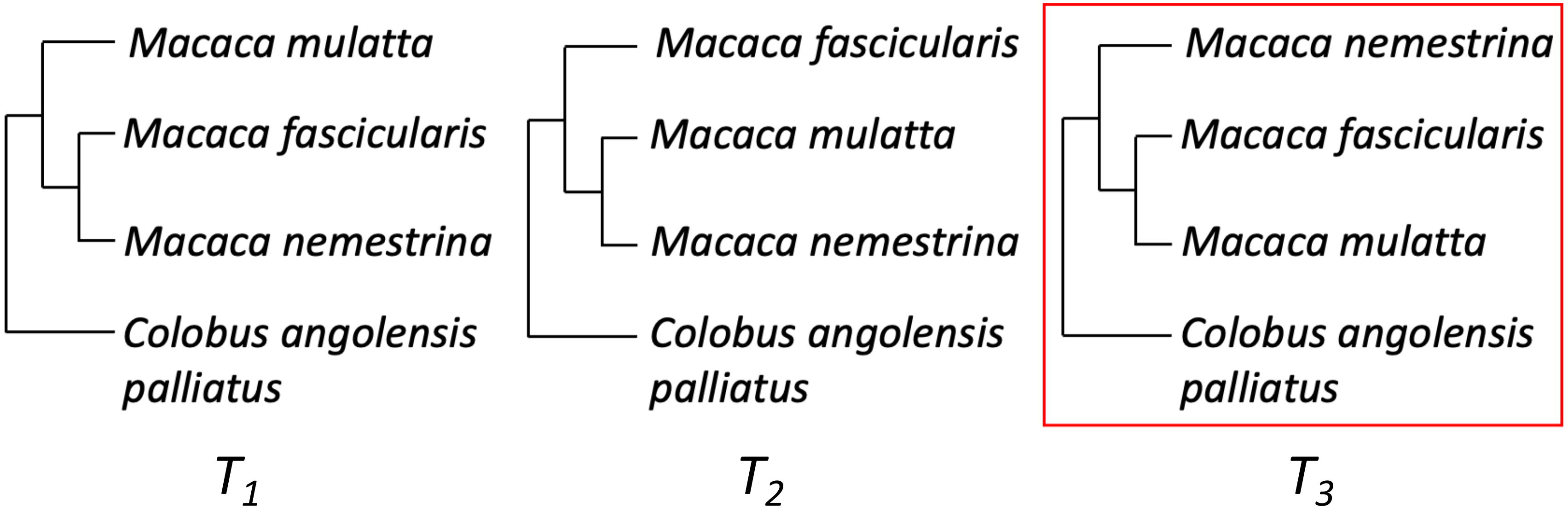
➢The frequencies of the minor trees from the MAST analysis are very similar

➢Good indication of the existence of incomplete lineage sorting in this data set

Number of genes: 1,599

Total length: 1,629,163

Apply to Macaques data

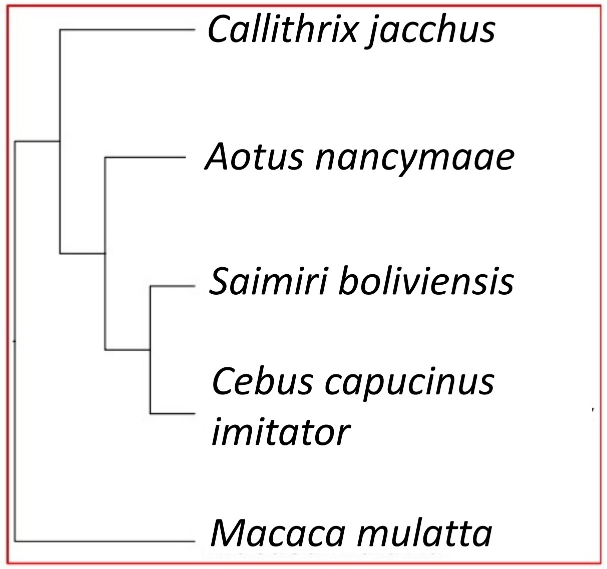


Gene tree frequencies:	31.2%	18.6%	50.2%
Best-fit MAST model weights:	**17.3%**	**14.2%**	**68.5%**

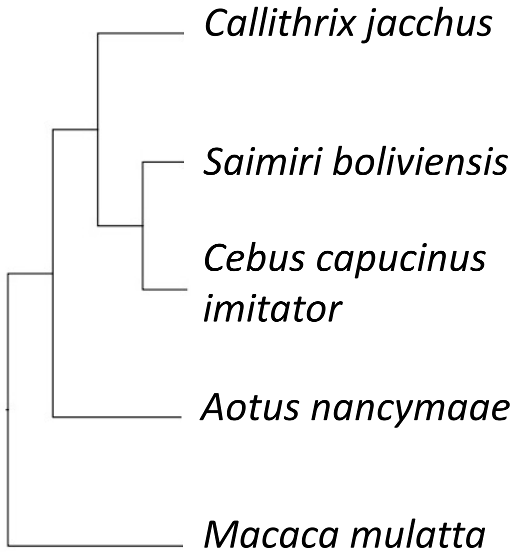
- ➢The minor trees are substantially different in frequency from the MAST analysis
- ➢Good indication of the existence of introgression in this data set

Apply to New World Monkeys

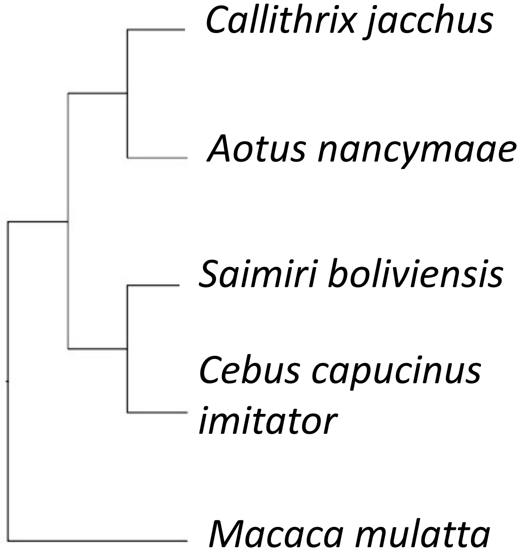
Number of genes: 1,557
Total length: 1,610,755



T_1



T_2



T_3

Gene tree frequencies:	37.1%	32.4%	30.5%
Single-tree model (IQ-TREE):			100%
Our result			
Best-fit MAST model weights:	**42.4%**	**28.1%**	**29.6%**

- ➢ The MAST model reported T_{D1} as the topology with the highest weight
- ➢ The MAST model can analyse a concatenated alignment using maximum likelihood, but without the biases that come with the single-tree assumption

Conclusion

1. The MAST model can overcome biases for a maximum likelihood method with single-tree assumption.

2. The weights estimated by the MAST model can be a good indication of some biological processes, like incomplete lineage sorting or introgression.

Published in Systematic Biology in March 2024 (https://doi.org/10.1093/sysbio/syae008)

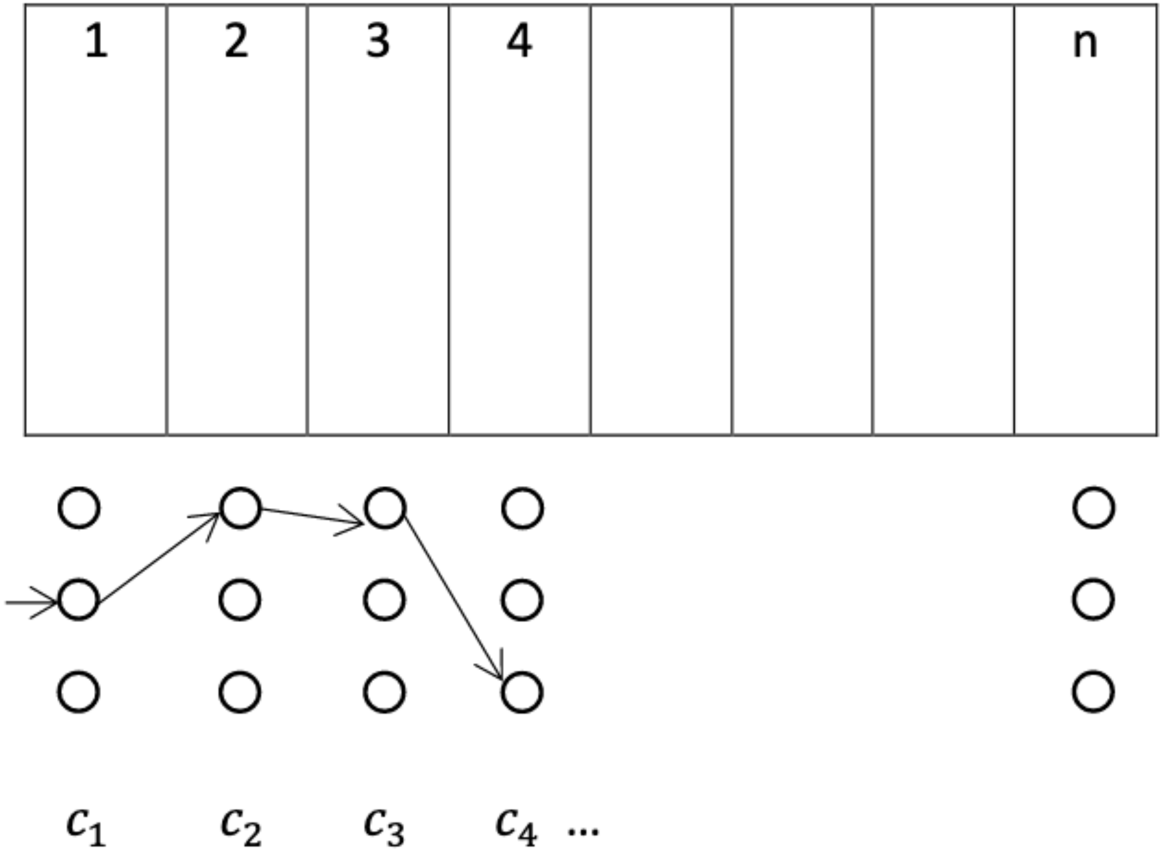
HMM-MAST model, considering dependence between topologies along the adjacent sites (Preliminary results)

HMM-MAST: HMM on multi-tree model

- MAST model does not consider the dependence between the sites.
- Topologies along the sites should have dependence.
- The model is based on the paper:

  J Felsenstein, G A Churchill, A Hidden Markov Model approach to variation among sites in rate of evolution., *Molecular Biology and Evolution*, Volume 13, Issue 1, Jan 1996, Pages 93–104, https://doi.org/10.1093/oxfordjournals.molbev.a025575

HMM-MAST: HMM on multi-tree model



- Assume each site has evolved under any of three trees {T_1, T_2, T_3}.
- However, this information is hidden.
- For each site, define the state c_i as the trees the site *i* may belong to.

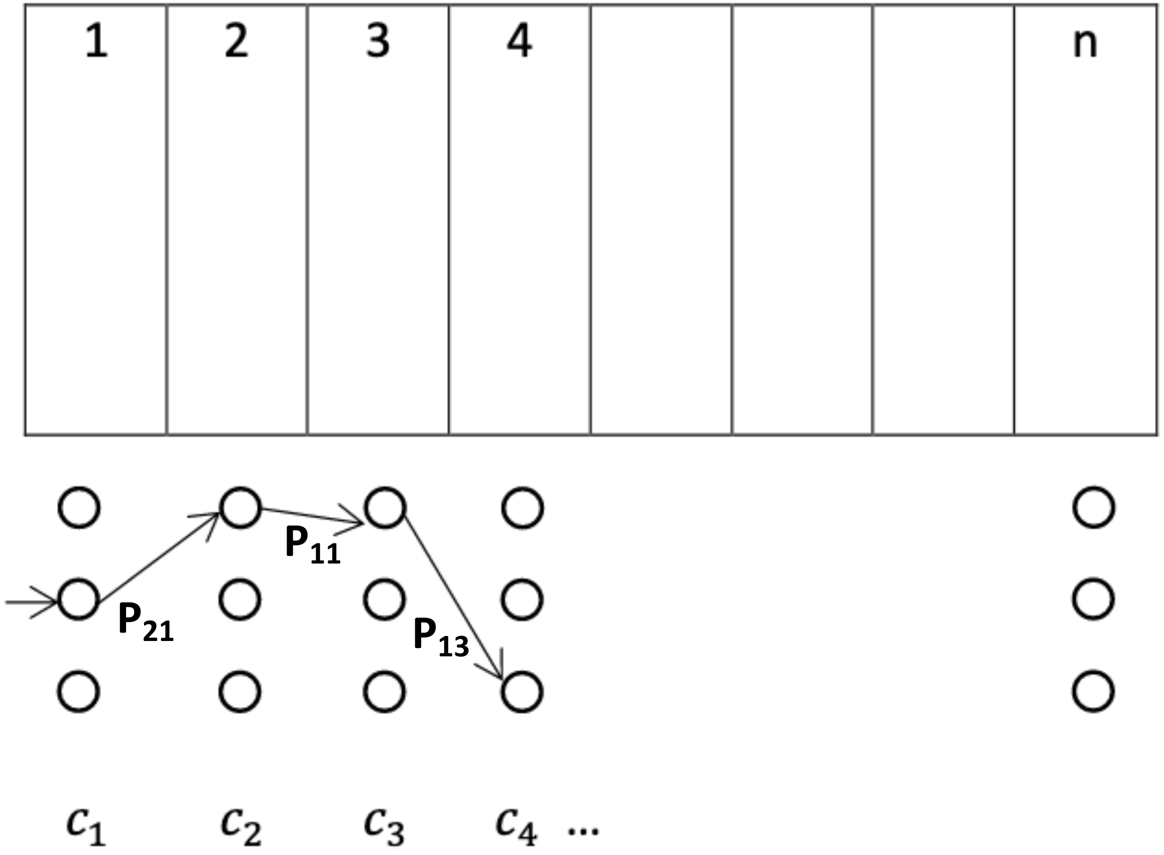
Parameters of the HMM-MAST

Transition probability matrix :

- Transition probability of going from c_{i-1} to c_i

$$P_{3x3} = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

The transition probability between the trees along the sites.



Backward probability formula

$$\Pr(D|T_1, T_2, T_3) = \sum_{c_1} \sum_{c_2} \dots \sum_{c_n} \Pr(c_1, c_2, \dots, c_n) \Pr\left(D \middle| T_{c_1}, T_{c_2}, \dots, T_{c_n}\right)$$

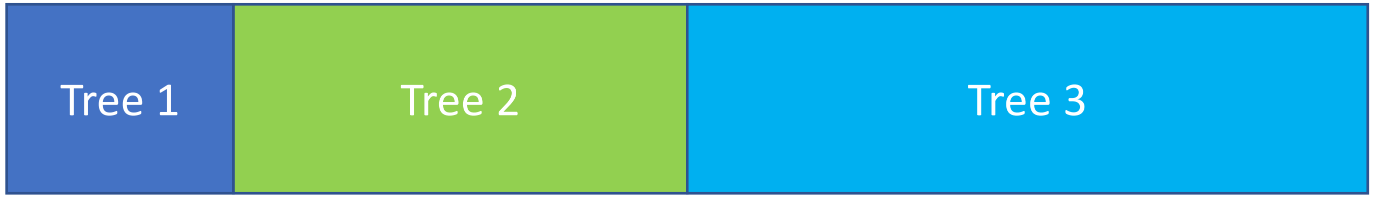
$$= \sum_{c_1=1}^{k} f_{c_1} Pr\left\{D_1 | T_{c_1}\right\} \sum_{c_2=1}^{k} p_{c_1 c_2} Pr\left\{D_2 | T_{c_2}\right\} \sum_{c_4=1}^{k} \dots \sum_{c_n=1}^{k} p_{c_{n-1} c_n} Pr\left\{D_n | T_{c_n}\right\}$$

Difference between our method and PhyML-multi

- PhyML-multi also implemented HMM on phylogenetic.
- PhyML-multi assumes each site evolves under each topology with equal probability.
- PhyML-multi has difficulty in handling long (> 5K) alignments

Simulated data sets with partitions

- Various numbers of partitions with different numbers of sequences
- Each partition was simulated using a different tree
- The model used to simulate the data sets: GTR+G
- Every trees have different GTR and Gamma models
- For example:

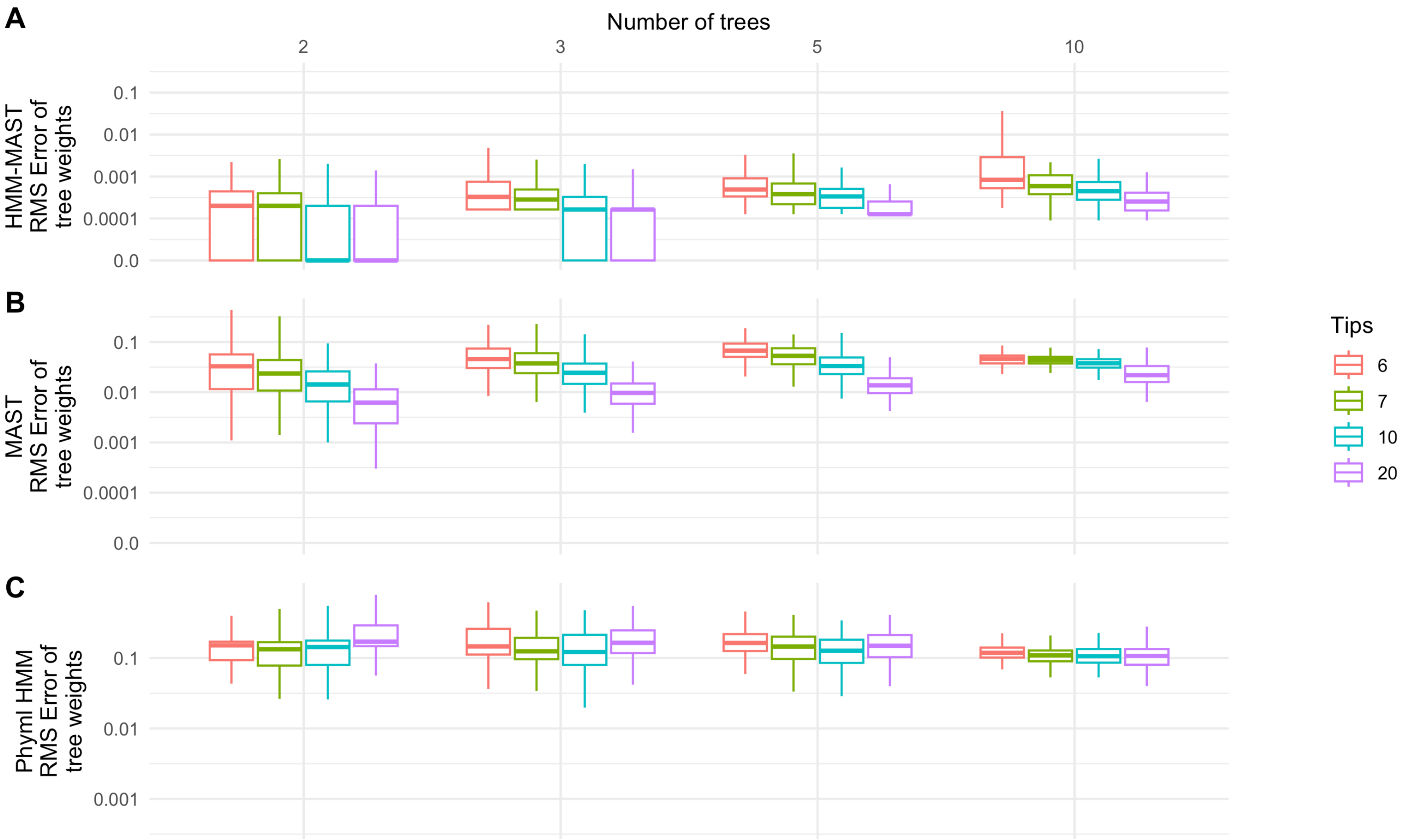


- Alignment length: 5K

Simulation results

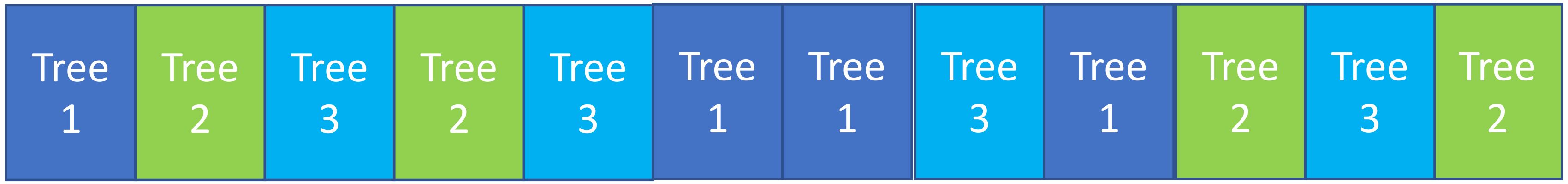
Evaluate the proportion of sites belonging each tree

Alignment length: 5K



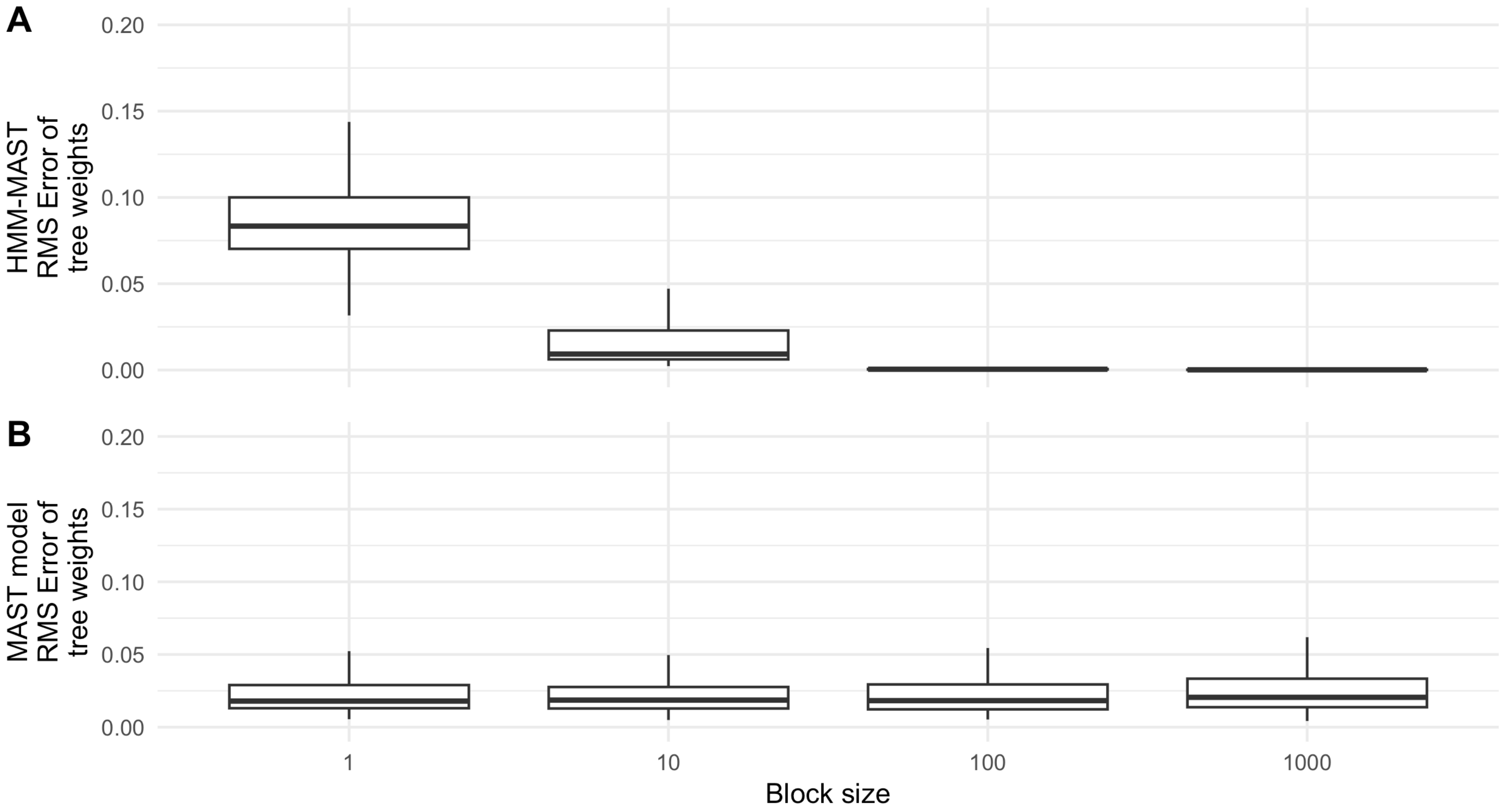
Simulated data sets with very short blocks

- Number of trees = 10
- Number of tips = 10
- Alignment length = 50K
- Block lengths: 1, 10, 100, 1000
- Each block was randomly assigned one tree
- The model used to simulate the data sets: GTR+G
- Every tree has different GTR and Gamma models
- For example (for 3 trees):



Tree 1	Tree 2	Tree 3	Tree 2	Tree 3	Tree 1	Tree 1	Tree 3	Tree 1	Tree 2	Tree 3	Tree 2

Results on simulated data sets with short blocks



Alignment lengths: 50K; Number of input trees = 10

Next direction…….

- HMM-MAST works well in simulated data sets
- Will perform testing on empirical data sets

Acknowledgements

Collaborators:

Indiana University, USA

- Matthew W Hahn

University of Auckland, New Zealand

- Allen G Rodrigo

Australian National University

- Minh Bui
- Rob Lanfear
- Caitlin Cherryh
- Huaiyan Ren
- Jeremias Ivan
- Nhan Trong Ly
- Rahil Vora